

Can AI-Generated Text be Reliably Detected?

Year: 2023 | Citations: 475 | Authors: Vinu Sankar Sadasivan, Aounon Kumar, S. Balasubramanian, Wenxiao Wang, S. Feizi

Abstract

Large Language Models (LLMs) perform impressively well in various applications. However, the potential for misuse of these models in activities such as plagiarism, generating fake news, and spamming has raised concern about their responsible use. Consequently, the reliable detection of AI-generated text has become a critical area of research. AI text detectors have shown to be effective under their specific settings. In this paper, we stress-test the robustness of these AI text detectors in the presence of an attacker. We introduce recursive paraphrasing attack to stress test a wide range of detection schemes, including the ones using the watermarking as well as neural network-based detectors, zero shot classifiers, and retrieval-based detectors. Our experiments conducted on passages, each approximately 300 tokens long, reveal the varying sensitivities of these detectors to our attacks. Our findings indicate that while our recursive paraphrasing method can significantly reduce detection rates, it only slightly degrades text quality in many cases, highlighting potential vulnerabilities in current detection systems in the presence of an attacker. Additionally, we investigate the susceptibility of watermarked LLMs to spoofing attacks aimed at misclassifying human-written text as AI-generated. We demonstrate that an attacker can infer hidden AI text signatures without white-box access to the detection method, potentially leading to reputational risks for LLM developers. Finally, we provide a theoretical framework connecting the AUROC of the best possible detector to the Total Variation distance between human and AI text distributions. This analysis offers insights into the fundamental challenges of reliable detection as language models continue to advance. Our code is publicly available at <https://github.com/vinusankars/Reliability-of-AI-text-detectors>.